政治学方法論 I

4. 統計分析の不確実性

矢内 勇生

大学院法学研究科・法学部

2016 年 5 月 11 日

神戸大学

今日の内容

1. ベイズ統計学?
   - ベイズの定理
   - ベイズ統計学とベイズの定理
2. 統計分析の不確実性
   - 統計的推定における不確実性の表現
   - ベイズ的な推論へ

ベイズの定理

- 推定の対象である母数 θ
- データ D

ベイズの定理 (Bayes Theorem)

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}$$

- $p(D|\theta) = L(\theta|D)$：尤度 (likelihood)
- $p(\theta)$：事前確率（事前分布, prior）
- $p(\theta|D)$：事後確率（事後分布, posterior）

ベイズの定理の証明

$p(D) \neq 0$, $p(\theta) \neq 0$ だとすると、

$$p(\theta|D) = \frac{p(\theta,D)}{p(D)},$$
$$p(D|\theta) = \frac{p(\theta,D)}{p(\theta)}$$

だから、

$$p(\theta,D) = p(\theta|D)p(D) = p(D|\theta)p(\theta)$$

である。したがって、

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

となり、ベイズの定理が導かれる。

ベイズの定理の使い方：よくある例

病気の診断

- ある病気の人 (sick: S) に対して実施すると 95%の確率で病気を見つける（陽性 [positive] と判定する）検査がある
- その病気ではない人 (healthy: H) に実施しても、1%の確率で陽性であると判定してしまう
- この病気に罹っているのは、人口の 0.1%である
- あなたがこの検査を受けたら「陽性」が出た
- 問題：あなたがこの病気に罹っている確率は？

ベイズの定理の使い方：よくある例

病気の診断

- ある病気の人 (sick: S) に対して実施すると 95%の確率で病気を見つける（陽性 [positive] と判定する）検査がある
- その病気ではない人 (healthy: H) に実施しても、1%の確率で陽性であると判定してしまう
- この病気に罹っているのは、人口の 0.1%である
- あなたがこの検査を受けたら「陽性」が出た
- 問題：あなたがこの病気に罹っている確率は？

直感的に、病気である確率は「高い」か「低い」か？

ベイズの定理を使って解く

- 知りたい確率：検査で陽性 (P) だったとき、病気 (S) である確率

$$\Pr(S|P)$$

ベイズの定理を使って解く

- 知りたい確率：検査で陽性 (P) だったとき、病気 (S) である確率

$$\Pr(S|P) = \frac{\Pr(P|S)\Pr(S)}{\Pr(P)} = \frac{\Pr(P|S)\Pr(S)}{\Pr(P|S)\Pr(S) + \Pr(P|H)\Pr(H)}$$

ベイズの定理を使って解く

- 知りたい確率：検査で陽性 (P) だったとき、病気 (S) である確率

$$\Pr(S|P) = \frac{\Pr(P|S)\Pr(S)}{\Pr(P)} = \frac{\Pr(P|S)\Pr(S)}{\Pr(P|S)\Pr(S) + \Pr(P|H)\Pr(H)}$$

- $\Pr(P|S) = 0.95$
- $\Pr(S) = 0.001$
- $\Pr(P|H) = 0.01$
- $\Pr(H) = 1 - \Pr(S) = 0.999$

ベイズの定理を使って解く

- 知りたい確率：検査で陽性 (P) だったとき、病気 (S) である確率

$$\Pr(S|P) = \frac{\Pr(P|S)\Pr(S)}{\Pr(P)} = \frac{\Pr(P|S)\Pr(S)}{\Pr(P|S)\Pr(S) + \Pr(P|H)\Pr(H)}$$

- $\Pr(P|S) = 0.95$
- $\Pr(S) = 0.001$
- $\Pr(P|H) = 0.01$
- $\Pr(H) = 1 - \Pr(S) = 0.999$

$$\Pr(S|P) = \frac{0.95 * 0.001}{0.95 * 0.001 + 0.01 * 0.999} \approx 0.0868$$

ベイズの定理でうまく解けるが...

- 直感に反し (?)、精度の高い検査で陽性でも、病気である確率は低い
- 理由：病気である（事前）確率が、ものすごく小さい
- 実際、陽性という結果は、病気である（事後）確率を上げている

ベイズの定理でうまく解けるが...

- 直感に反し (?)、精度の高い検査で陽性でも、病気である確率は低い
- 理由：病気である（事前）確率が、ものすごく小さい
- 実際、陽性という結果は、病気である（事後）確率を上げている
- 問題
  - ベイズの定理がないとこの問題は解けないか?
  - ベイズの定理を使えば、「ベイジアン」なのか？

問題を捉え直す

- **無作為に選ばれた 10 万人がこの検査を受けると仮定する**
- 10 万人のうち、100 人がこの病気に罹っており、残りの 99,900 人はこの病気ではない
- 病気の 100 人のうち、95 人が陽性、5 人が陰性という結果を得る
- 健康な 99,900 人のうち、999 人が陽性、残りは陰性という結果を得る
- 問題：陽性という結果を受けた人のうち、病気の人の割合は？

問題を捉え直す

- **無作為に選ばれた 10 万人がこの検査を受けると仮定する**
- 10 万人のうち、100 人がこの病気に罹っており、残りの 99,900 人はこの病気ではない
- 病気の 100 人のうち、95 人が陽性、5 人が陰性という結果を得る
- 健康な 99,900 人のうち、999 人が陽性、残りは陰性という結果を得る
- 問題：陽性という結果を受けた人のうち、病気の人の割合は？

$$\Pr(S|P) = \frac{95}{95+999} \approx 0.0868$$

ベイズと頻度論

- 同じ問題でも、頻度で捉えた方が自然に解決できることがある
- ベイズの定理をあえて使わなくても、頻度を順番に考えれば、正しい答えに到達できる

ベイズと頻度論

- 同じ問題でも、頻度で捉えた方が自然に解決できることがある
- ベイズの定理をあえて使わなくても、頻度を順番に考えれば、正しい答えに到達できる
- ベイズの定理を使うことが、ベイジアンの目的ではない
  1. 頻度を数えていても、ベイジアンかもしれない
  2. ベイズの定理を使っていても、頻度主義者かもしれない
- 話をわかりやすくする１つの方法：数を増やす
- 病気の検査の例
  - １人に対する検査：ベイズの定理が必要
  - 10 万人に対する検査：頻度の比で割合（確率）がわかる

ベイズと頻度論

- 同じ問題でも、頻度で捉えた方が自然に解決できることがある
- ベイズの定理をあえて使わなくても、頻度を順番に考えれば、正しい答えに到達できる
- ベイズの定理を使うことが、ベイジアンの目的ではない
  1. 頻度を数えていても、ベイジアンかもしれない
  2. ベイズの定理を使っていても、頻度主義者かもしれない
- 話をわかりやすくする 1 つの方法：数を増やす
- 病気の検査の例
  - 1 人に対する検査：ベイズの定理が必要
  - 10 万人に対する検査：頻度の比で割合（確率）がわかる
- コンピュータによるサンプリング（シミュレーション）を行う！

点推定と区間推定

例題：コイン投げの枚数を推定する

表が出る確率が 0.5 であるコインを N 枚投げたところ、表が 10 枚出た。投げたコインの枚数 N はいくつか？

点推定と区間推定

例題：コイン投げの枚数を推定する

表が出る確率が 0.5 であるコインを N 枚投げたところ、表が 10 枚出た。投げたコインの枚数 N はいくつか？

- 点推定：$N = 20$
- 区間推定：$N \in \{13, 14, \ldots, 30\}$

点推定と区間推定

例題：コイン投げの枚数を推定する

表が出る確率が 0.5 であるコインを N 枚投げたところ、表が 10 枚出た。投げたコインの枚数 N はいくつか？

- 点推定：$N = 20$
- 区間推定：$N \in \{13, 14, \ldots, 30\}$

- 点推定
  - ピンポイントな推定
  - 推定がどれくらい信頼できるか不明
- 区間推定
  - 推定結果に幅がある：1 つの推定値を確実視しない
  - 区間の幅によって、不確実性を表現できる

信頼区間 (confidence intervals)

- 自由度 $n-k$ の t 分布の上側 100α パーセンタイルを $t_{n-k}(\alpha)$ とする
- 母数 θ の点推定値 $\hat{\theta}$
- 母数 θ の $100(1-\alpha)$ パーセント信頼区間：

$$[\hat{\theta} - t_{n-k}(\alpha/2) \cdot \mathrm{se},\ \hat{\theta} + t_{n-k}(\alpha/2) \cdot \mathrm{se}]$$

信頼区間 (confidence intervals)

- 自由度 $n-k$ の t 分布の上側 100α パーセンタイルを $t_{n-k}(\alpha)$ とする
- 母数 θ の点推定値 $\hat{\theta}$
- 母数 θ の $100(1-\alpha)$ パーセント信頼区間：

$$[\hat{\theta} - t_{n-k}(\alpha/2) \cdot \mathrm{se},\ \hat{\theta} + t_{n-k}(\alpha/2) \cdot \mathrm{se}]$$

- 95 パーセント信頼区間：同じ母集団から、同じ手続きでデータを抽出して分析するという作業を繰り返し行ったとき、求めた信頼区間のうちの 95%は母数の真の値を区間内に含む
  - 「この信頼区間に母数が含まれる確率が 95%」ではない!
  - 1 つの信頼区間に母数が含まれる確率は、0 か 1 のいずれか

非ベイズ的な仮説検定

- 理論：ある説明変数 X が結果変数 Y になんらかの影響 β を与える
- 統計分析のための仮説

  帰無仮説： $\beta = 0$
  対立仮説： $\beta \neq 0$
- 目標：帰無仮説が正しくないことを示し、効果が 0 でないと結論する
- 必要な手続き
  1. 統計量を計算する
  2. 帰無仮説が正しいと仮定し、上の統計量が得られる確率を計算する
  3. 確率が著しく小さいとき、仮定が誤りであると判断し、帰無仮説を棄却する

p 値 (p values)

- 説明変数が結果変数に影響を与えていない（$\beta = 0$、すなわち帰無仮説が正しい）ときに、現在分析中のデータまたはより極端なデータを得る確率
  - 「帰無仮説が正しい確率」ではない!
  - 「対立仮説が間違っている確率」ではない!
- p 値（帰無仮説が正しいと仮定して計算）が小さい
  $\rightarrow$ 分析中のデータを得る確率は小さい（にも拘らず、現にデータを持っている）
  $\rightarrow$ 帰無仮説が間違っていると考えることにする（帰無仮説を棄却する）
- 「p 値 = 有意水準」ではない!
- p 値はデータから計算するもの、有意水準は自分で（恣意的に）決めるもの

検定と統計的有意性 (statistical significance)

有意水準を 0.05 に設定すると

- $p < 0.05$ なら帰無仮説が棄却される
- おおよそ $\hat{\beta} \pm 2\text{se}$ の範囲に 0 が含まれないとき、その係数をもつ説明変数の効果の向き（正か負か）がはっきりする
- そのとき、その効果は「統計的に有意である」とされる
- 統計的有意は、効果の向きをはっきりさせるだけで、効果の大きさについては何も示さない
- 実際の研究では効果の大きさ（実質的に意味があるのか, substantive significance）を示すことが必要かつ重要
- 有意水準を 0.05 にしなければいけない理論的論拠はない

$\rightarrow$ 「$p < 0.05$ だから良い結果」ではない！！！

事後確率（事後分布）を使って推論する

- 点推定値として、事後分布の中で確率密度が最も高い点を使う：最大事後確率 (Maximum a posteriori: **MAP**)
- 区間推定：事後分布の中で、特定の確率を構成する区間を求める（詳しくは web 資料を参照）
  - 事後確率が高い部分から順に足しわせて構成した区間：最高事後密度区間 (highest posterior density interval: **HPDI** または HDI)
- 母数 β の 77%HPDI：β がその区間に入る確率が 77%（自然な解釈！）
- 母数が実質的に意味を持つ確率を計算することができる
  - 例えば、$\beta > 1$ でないと実質的には効果があるとは言えないなら、事後分布を使って $\beta > 1$ になる確率を求めればよい